ISIT'2013 1569719797

Fixed Points of Generalized Approximate Message Passing with Arbitrary Matrices

Sundeep Rangan*, Philip Schniter†, Erwin Riegler‡, Alyson Fletcher§, and Volkan Cevher¶
*NYU-Poly, Elec. & Comp. Engineering, srangan@poly.edu
†The Ohio State Univ., Elec. & Comp. Engineering, schniter@ece.osu.edu
‡Technische Institüt Wien, Inst. for Telecommunications, erwin.riegler@tuwien.ac.at
§UC Santa Cruz, Elec. & Comp. Engineering, afletcher@ucsc.edu
¶EPFL, Electrical Engineering, volkan.cevher@epfl.ch

***Abstract*—The estimation of a random vector with independent components passed through a linear transform followed by a componentwise (possibly nonlinear) output map arises in a range of applications. Approximate message passing (AMP) methods, based on Gaussian approximations of loopy belief propagation, have recently attracted considerable attention for such problems. For large random transforms, these methods exhibit fast convergence and admit precise analytic characterizations with testable conditions for optimality, even for certain non-convex problem instances. However, the behavior of AMP under general transforms is not fully understood. In this paper, we consider the generalized AMP (GAMP) algorithm and relate the method to more common optimization techniques. This analysis enables a precise characterization of the GAMP algorithm fixed-points that applies to arbitrary transforms. In particular, we show that the fixed points of the so-called max-sum GAMP algorithm for MAP estimation are critical points of a constrained maximization of the posterior density. The fixed-points of the sum-product GAMP algorithm for estimation of the posterior marginals can be interpreted as critical points of a certain mean-field variational optimization.**

***Index Terms*—Belief propagation, ADMM, variational optimization, message passing.**

I. Introduction

Consider the constrained optimization problem

$$(\widehat{\mathbf{x}},\widehat{\mathbf{z}}) := \arg\min_{\mathbf{x},\mathbf{z}} F(\mathbf{x},\mathbf{z}) \quad \text{s.t. } \mathbf{z}=\mathbf{A}\mathbf{x}, \tag{1}$$

where $\mathbf{x}\in\mathbb{R}^n$, $\mathbf{z}\in\mathbb{R}^m$, $\mathbf{A}\in\mathbb{R}^{m\times n}$ and the objective function admits a decomposition of the form

$$\begin{aligned} F(\mathbf{x},\mathbf{z}) &:= f_x(\mathbf{x}) + f_z(\mathbf{z}) \\ f_x(\mathbf{x}) = \sum_{j=1}^{n} f_{x_j}(x_j), &\quad f_z(\mathbf{z}) = \sum_{i=1}^{m} f_{z_i}(z_i), \end{aligned} \tag{2}$$

for scalar functions $f_{x_j}(\cdot)$ and $f_{z_i}(\cdot)$. One example where this optimization arises is the estimation problem in Fig. 1. Here, a random vector $\mathbf{x}$ has independent components with densities $p_{x_j}(x_j)$, and passes through a linear transform to yield an output $\mathbf{z}=\mathbf{A}\mathbf{x}$. The problem is to estimate $\mathbf{x}$ and $\mathbf{z}$ from measurements $\mathbf{y}$ generated by a componentwise conditional density $p_{y_i|z_i}(y_i|z_i)$. Under this observation model, the vectors $\mathbf{x}$ and $\mathbf{z}$ will have a posterior joint density given by

$$p_{\mathbf{x},\mathbf{z}|\mathbf{y}}(\mathbf{x},\mathbf{z}|\mathbf{y}) = [Z(\mathbf{y})]^{-1} e^{-F(\mathbf{x},\mathbf{z})}\, \mathbb{1}_{\{\mathbf{z}=\mathbf{A}\mathbf{x}\}}, \tag{3}$$

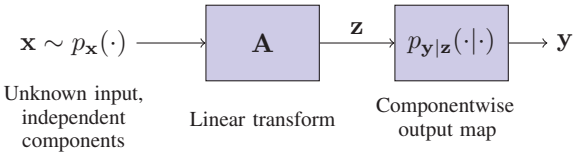

Fig. 1. System model: The GAMP method considered here can be used for approximate MAP and MMSE estimation of $\mathbf{x}$ from $\mathbf{y}$.

where $F(\mathbf{x},\mathbf{z})$ is given by (2) when the scalar functions are set to the negative log density and likelihood:

$$f_{x_j}(x_j) = -\log p_{x_j}(x_j), \quad f_{z_i}(z_i) = -\log p_{y_i|z_i}(y_i|z_i).$$

Note that in (3), $F(\mathbf{x},\mathbf{z})$ is implicitly a function of $\mathbf{y}$, $Z(\mathbf{y})$ is a normalization constant and the point mass $\mathbb{1}_{\{\mathbf{z}=\mathbf{A}\mathbf{x}\}}$ imposes the linear constraint that $\mathbf{z}=\mathbf{A}\mathbf{x}$. The optimization (1) in this case produces the *maximum a posteriori* (MAP) estimate of $\mathbf{x}$ and $\mathbf{z}$. In statistics, the system in Fig. 1 is sometimes referred to as a generalized linear model [1] and is used in a range of applications including regression, inverse problems, and filtering. Bayesian forms of compressed sensing can also be considered in this framework by imposing a sparse prior for the components x_j [2]. In all these applications, one may instead be interested in estimating the posterior marginals $p(x_j|\mathbf{y})$ and $p(z_i|\mathbf{y})$. We relate this objective to an optimization of the form (1)–(2) in the sequel.

Most current numerical methods for solving the constrained optimization problem (1) attempt to exploit the separable structure of the objective function (2) either through generalizations of the iterative shrinkage and thresholding (ISTA) algorithms [3]–[8] or alternating direction method of multipliers (ADMM) approach [9]–[12]. There are now a large number of these methods and we provide a brief review in Section II.

However, in recent years, there has been considerable interest in so-called approximate message passing (AMP) methods based on Gaussian and quadratic approximations of loopy belief propagation in graphical models [13]–[18]. The main appealing feature of the AMP algorithms is that for certain large random matrices $\mathbf{A}$, the asymptotic behavior of the algorithm can be rigorously and exactly predicted with testable conditions for optimality, even for many non-convex

instances. Moreover, in the case of these large, random matrices, simulations appear to show very fast convergence of AMP methods when compared against state-of-the-art conventional optimization techniques.

However, the behavior of AMP methods under general $\mathbf{A}$ is not fully understood. The broad purpose of this paper is to show that certain forms of AMP algorithms can be seen as variants of more conventional optimization methods. This analysis will enable a precise characterization of the fixed points of the AMP methods that applies to arbitrary $\mathbf{A}$.

Our study focuses on a generalized AMP (GAMP) method proposed in [18] and rigorously analyzed in [19]. We consider this algorithm since many other variants of AMP are special cases of this general procedure. The GAMP method has two common versions: max-sum GAMP for the MAP estimation of the vectors $\mathbf{x}$ and $\mathbf{z}$ for the problem in Fig. 1; and sum-product GAMP for approximate inference of the posterior marginals.

For both versions of GAMP, the algorithms produce estimates $\mathbf{x}$ and $\mathbf{z}$ along with certain "quadratic" terms. Our first main result (Theorem 1) shows that the fixed points $(\widehat{\mathbf{x}}, \widehat{\mathbf{z}})$ of the max-sum GAMP are critical points of the optimization (1). In addition, the quadratic terms can be considered as diagonal approximations of the inverse Hessian of the objective function. For sum-product GAMP, we provide a variational interpretation of the algorithm's fixed points. Specifically, we show (Theorem 2) that the algorithm's fixed points can be interpreted as means and variances of a certain Gaussian mean-field approximation of the posterior distribution.

A full version of the paper [20] includes all the proofs and more detailed discussions.

Algorithm 1 Generalized Approximate Message Passing

Require: Matrix $\mathbf{A}$, functions $f_x(\mathbf{x})$, $f_z(\mathbf{z})$. and algorithm choice MaxSum or SumProduct.

1: $t \leftarrow 0$
2: Initialize $\mathbf{x}^t$, $\boldsymbol{\tau}_x^t$
3: $\mathbf{s}^{t-1} \leftarrow 0$
4: $\mathbf{S} \leftarrow |\mathbf{A}|^2$ (componentwise magnitude squared)
5: **repeat**
6: {Output node update}
7: $\boldsymbol{\tau}_p^t \leftarrow \mathbf{S}\boldsymbol{\tau}_x^t$
8: $\mathbf{p}^t \leftarrow \mathbf{A}\mathbf{x}^t - \mathbf{s}^{t-1}\boldsymbol{\tau}_p^t$
9: **if** MaxSum **then**
10: $\mathbf{z}^t \leftarrow \text{prox}_{\boldsymbol{\tau}_p^t f_z}(\mathbf{p}^t)$
11: $\boldsymbol{\tau}_z^t \leftarrow \boldsymbol{\tau}_p^t \text{prox}'_{\boldsymbol{\tau}_p^t f_z}(\mathbf{p}^t)$
12: **else if** SumProduct **then**
13: $\mathbf{z}^t \leftarrow \mathbb{E}(\mathbf{z}|\mathbf{p}^t, \boldsymbol{\tau}_p^t)$
14: $\boldsymbol{\tau}_z^t \leftarrow \text{var}(\mathbf{z}|\mathbf{p}^t, \boldsymbol{\tau}_p^t)$
15: **end if**
16: $\mathbf{s}^t \leftarrow (\mathbf{z}^t - \mathbf{p}^t)/\boldsymbol{\tau}_p^t$
17: $\boldsymbol{\tau}_s^t \leftarrow 1/\boldsymbol{\tau}_p^t - \boldsymbol{\tau}_z^t/(\boldsymbol{\tau}_p^t)^2$
18:
19: {Input node update}
20: $\boldsymbol{\tau}_r^t \leftarrow 1/(\mathbf{S}^T\boldsymbol{\tau}_s^t)$
21: $\mathbf{r}^t \leftarrow \mathbf{x}^t + \boldsymbol{\tau}_r^t\mathbf{A}^T\mathbf{s}^t$
22: **if** MaxSum **then**
23: $\mathbf{x}^{t+1} \leftarrow \text{prox}_{\boldsymbol{\tau}_r^t f_x}(\mathbf{r}^t)$
24: $\boldsymbol{\tau}_x^{t+1} \leftarrow \boldsymbol{\tau}_r^t \text{prox}'_{\boldsymbol{\tau}_r^t f_x}(\mathbf{r}^t)$
25: **else if** SumProduct **then**
26: $\mathbf{x}^{t+1} \leftarrow \mathbb{E}(\mathbf{x}|\mathbf{r}^t, \boldsymbol{\tau}_r^t)$
27: $\boldsymbol{\tau}_x^{t+1} \leftarrow \text{var}(\mathbf{x}|\mathbf{r}^t, \boldsymbol{\tau}_r^t)$
28: **end if**
29: **until** Terminated

II. Review of GAMP and Related Methods

A. Generalized Approximate Message Passing

Graphical-model methods [21] are a natural approach to the optimization problem (1) given the separable structure of the objective function (2). However, traditional graphical model techniques such as loopy belief propagation (loopy BP) generally require that the constraint matrix $\mathbf{A}$ is sparse. Approximate message passing (AMP) refers to a class of Gaussian and quadratic approximations of loopy BP that can be applied to dense $\mathbf{A}$. AMP approximations of loopy BP originated in CDMA multiuser detection problems [22] and have received considerable recent attention in the context of compressed sensing [13]–[18]. The Gaussian approximations used in AMP are also closely related to expectation propagation techniques [23].

In this work, we study the so-called generalized AMP (GAMP) algorithm [18] rigorously analyzed in [19]. The procedure, shown in Algorithm 1, produces a sequence of estimates $(\mathbf{x}^t, \mathbf{z}^t)$ along with the *quadratic terms* $\boldsymbol{\tau}_x^t, \boldsymbol{\tau}_z^t, \boldsymbol{\tau}_p^t, \boldsymbol{\tau}_r^t, \boldsymbol{\tau}_s^t$. Note that all vector-vector multiplications in Algorithm 1 (e.g., $\mathbf{s}^{t-1}\boldsymbol{\tau}_p^t$) are to be taken componentwise.

We focus on two variants of the GAMP algorithm: *max-sum GAMP* and *sum-product GAMP*. In the max-sum version of the algorithm, the outputs $(\mathbf{x}^t, \mathbf{z}^t)$ represent estimates of the solution to the optimization problem (1), or equivalently the MAP estimates for the posterior (3). Since the objective function has the separable form (2), each iteration of the algorithm involves four componentwise update steps: the proximal updates shown in lines 10 and 23, where

$$\text{prox}_f(v) := \arg\min_{u \in \mathbb{R}} f(u) + \frac{1}{2}|u - v|^2, \tag{4}$$

and lines 11 and 24, involving the derivative of the proximal operator from (4). Thus, max-sum GAMP reduces the vector-valued optimization (1) to a sequence of scalar optimizations.

For the sum-product GAMP algorithm, the outputs $(\mathbf{x}^t, \mathbf{z}^t)$ represent the posterior means for the density (3), or equivalently, the minimum mean-squared error (MMSE) estimates of $(\mathbf{x}, \mathbf{z})$ given $\mathbf{y}$. As discussed in [18], the algorithm also provides estimates of the posterior marginals. The expectations and variances in lines 13, 14, 26 and 27 of the algorithm are taken with respect to the probability density functions:

$$p(\mathbf{x}|\mathbf{r}, \boldsymbol{\tau}_r) \propto \exp\left[-f_x(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{r}\|^2_{\boldsymbol{\tau}_r}\right] \tag{5a}$$

$$p(\mathbf{z}|\mathbf{p}, \boldsymbol{\tau}_p) \propto \exp\left[-f_z(\mathbf{z}) + \frac{1}{2}\|\mathbf{z} - \mathbf{p}\|^2_{\boldsymbol{\tau}_p}\right], \tag{5b}$$

where, for any vectors $\mathbf{v} \in \mathbb{R}^r$ and $\boldsymbol{\tau} \in \mathbb{R}^r$ with $\boldsymbol{\tau} > 0$,

$$\|\mathbf{v}\|_{\boldsymbol{\tau}}^2 := \sum_{i=1}^{r} \frac{|v_i|^2}{\tau_i}.$$

Under separability assumption (2), these densities factor as

$$p(\mathbf{x}|\mathbf{r}, \boldsymbol{\tau}_r) \propto \prod_{j=1}^{n} \exp\left[-f_{x_j}(x_j) - \frac{|x_j - r_j|^2}{2\tau_{r_j}}\right] \quad (6a)$$

$$p(\mathbf{z}|\mathbf{p}, \boldsymbol{\tau}_p) \propto \prod_{i=1}^{m} \exp\left[-f_{z_i}(z_i) - \frac{|z_i - p_i|^2}{2\tau_{p_i}}\right], \quad (6b)$$

allowing the expectation and variance computations to be computed componentwise, and reducing the sum-product GAMP algorithm to a sequence of scalar estimation optimizations.

B. Iterative Shrinkage and Thresholding Algorithm

The goal in the paper is to relate the GAMP method to more conventional optimization techniques. One of the more common of such approaches is a generalization of the Iterative Shrinkage and Thresholding Algorithm (ISTA) shown in Algorithm 2 [3]–[6].

Algorithm 2 Iterative Shrinkage and Thresholding Algorithm

Require: Matrix $\mathbf{A}$, scalar c, functions $f_x(\cdot)$, $f_z(\cdot)$.
1: $t \leftarrow 0$
2: Initialize $\mathbf{x}^t$.
3: **repeat**
4: $\quad \mathbf{z}^t \leftarrow \mathbf{A}\mathbf{x}^t$
5: $\quad \mathbf{q}^t \leftarrow \partial f_z(\mathbf{z}^t)/\partial \mathbf{z}$
6: $\quad \mathbf{x}^{t+1} \leftarrow \arg\min_{\mathbf{x}} f_x(\mathbf{x}) + (\mathbf{q}^t)^T\mathbf{A}\mathbf{x} + (c/2)\|\mathbf{x} - \mathbf{x}^t\|^2$
7: **until** Terminated

The algorithm is built on the idea that, at each iteration t, the second cost term in the minimization $\arg\min_{\mathbf{x}} f_x(\mathbf{x}) + f_z(\mathbf{A}\mathbf{x})$ specified by (1) is replaced by a quadratic majorizing cost $g_z(\mathbf{x}) \geq f_z(\mathbf{A}\mathbf{x})$ that coincides at the point $\mathbf{x} = \mathbf{x}^t$ (i.e., $g_z(\mathbf{x}^t) = f_z(\mathbf{A}\mathbf{x}^t)$), where majorization can be achieved via appropriate choice of $c > 0$. This approach is motivated by the fact that, if $f_x(\mathbf{x})$ and $f_z(\mathbf{z})$ are both separable, as in (2), then both the gradient in line 5 and minimization in line 6 can be performed componentwise. Moreover, when $f_x(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$, as in the LASSO problem [24], the minimization in line 6 can be computed directly via a shrinkage and thresholding operation – hence the name of the algorithm. The convergence of the ISTA method tends to be slow, but a number of enhanced methods have been successful and widely-used [5]–[8].

C. Alternating Direction Method of Multipliers

A second common class of methods is built around the Alternating Direction Method of Multipliers (ADMM) [9] approach shown in Algorithm 3. The Lagrangian for the optimization problem (1) is given by

$$L(\mathbf{x}, \mathbf{z}, \mathbf{s}) := F(\mathbf{x}, \mathbf{z}) + \mathbf{s}^T(\mathbf{z} - \mathbf{A}\mathbf{x}), \quad (7)$$

where $\mathbf{s}$ are the dual parameters. ADMM attempts to produce a sequence of estimates $(\mathbf{x}^t, \mathbf{z}^t, \mathbf{s}^t)$ that converge to a saddle point of the Lagrangian (7). The parameters of the algorithm are a step-size $\alpha > 0$ and the penalty terms $Q_z(\cdot)$ and $Q_x(\cdot)$, which classical ADMM would choose as

$$Q_x(\mathbf{x}, \mathbf{x}^t, \mathbf{z}^t, \alpha) = \frac{\alpha}{2}\|\mathbf{z}^t - \mathbf{A}\mathbf{x}\|^2 \quad (8a)$$

$$Q_z(\mathbf{z}, \mathbf{z}^t, \mathbf{x}^{t+1}, \alpha) = \frac{\alpha}{2}\|\mathbf{z} - \mathbf{A}\mathbf{x}^{t+1}\|^2. \quad (8b)$$

Algorithm 3 Alternating Direction Method of Multipliers

Require: $\mathbf{A}$, α, functions $f_x(\cdot)$, $f_z(\cdot)$, $Q_x(\cdot)$, $Q_z(\cdot)$
1: $t \leftarrow 0$
2: Initialize $\mathbf{x}^t$, $\mathbf{z}^t$, $\mathbf{s}^t$
3: **repeat**
4: $\quad \mathbf{x}^{t+1} \leftarrow \arg\min_{\mathbf{x}} L(\mathbf{x}, \mathbf{z}^t, \mathbf{s}^t) + Q_x(\mathbf{x}, \mathbf{x}^t, \mathbf{z}^t, \alpha)$
5: $\quad \mathbf{z}^{t+1} \leftarrow \arg\min_{\mathbf{z}} L(\mathbf{x}^{t+1}, \mathbf{z}, \mathbf{s}^t) + Q_z(\mathbf{z}, \mathbf{z}^t, \mathbf{x}^{t+1}, \alpha)$
6: $\quad \mathbf{s}^{t+1} \leftarrow \mathbf{s}^t + \alpha(\mathbf{z}^{t+1} - \mathbf{A}\mathbf{x}^{t+1})$
7: **until** Terminated

When the objective function admits a separable form (2) and one uses the auxiliary function $Q_z(\cdot)$ in (8b), the $\mathbf{z}$-minimization in line 5 separates into m scalar optimizations. However, due to the quadratic term $\|\mathbf{A}\mathbf{x}\|^2$ in (8a), the $\mathbf{x}$-minimization in line 4 does not separate for general $\mathbf{A}$. To circumvent this problem, one might consider a separable inexact $\mathbf{x}$-minimization, since many inexact variants of ADMM are known to converge [9]. For example, $Q_x(\cdot)$ might be chosen to yield separability while majorizing the original cost in line 4, as was done for ISTA's line 6. Many other choices of penalty $Q_x(\cdot)$ have also been considered in the literature (see, e.g., the overview in [10]).

Other variants of ADMM are also possible [9]. For example, the step-size α might vary with the iteration t, or the penalty terms might have the form $(\mathbf{z} - \mathbf{A}\mathbf{x})^T\mathbf{P}(\mathbf{z} - \mathbf{A}\mathbf{x})$ for positive semidefinite $\mathbf{P}$. As we will see, these generalizations provide a connection to GAMP.

III. Fixed-Points of Max-Sum GAMP

Our first result connects the max-sum GAMP algorithm to inexact ADMM.

Theorem 1: The outputs of the max-sum GAMP version of Algorithm 1 satisfy the recursions

$$\mathbf{x}^{t+1} = \arg\min_{\mathbf{x}} \left[L(\mathbf{x}, \mathbf{z}^t, \mathbf{s}^t) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}^t\|_{\boldsymbol{\tau}_r^t}^2\right] \quad (9a)$$

$$\mathbf{z}^{t+1} = \arg\min_{\mathbf{z}} \left[L(\mathbf{x}^{t+1}, \mathbf{z}, \mathbf{s}^t) + \frac{1}{2}\|\mathbf{z} - \mathbf{A}\mathbf{x}^{t+1}\|_{\boldsymbol{\tau}_p^{t+1}}^2\right] \quad (9b)$$

$$\mathbf{s}^{t+1} = \mathbf{s}^t + \frac{1}{\boldsymbol{\tau}_p^{t+1}}(\mathbf{z}^{t+1} - \mathbf{A}\mathbf{x}^{t+1}) \quad (9c)$$

where $L(\mathbf{x}, \mathbf{z}, \mathbf{s})$ is the Lagrangian defined in (7).

Now suppose that $(\widehat{\mathbf{x}}, \widehat{\mathbf{z}}, \mathbf{s}, \boldsymbol{\tau}_x, \boldsymbol{\tau}_s)$ is a fixed point of the algorithm (where the "hats" on $\widehat{\mathbf{x}}$ and $\widehat{\mathbf{z}}$ are used to distinguish them from free variables). Then, this fixed point is a critical point of the constrained optimization (1) in that $\widehat{\mathbf{z}} = \mathbf{A}\widehat{\mathbf{x}}$ and

$$\frac{\partial}{\partial \mathbf{x}} L(\widehat{\mathbf{x}}, \widehat{\mathbf{z}}, \mathbf{s}) = \mathbf{0}, \quad \frac{\partial}{\partial \mathbf{z}} L(\widehat{\mathbf{x}}, \widehat{\mathbf{z}}, \mathbf{s}) = \mathbf{0}. \quad (10)$$

The first part of the Theorem, equation (9), shows that max-sum GAMP can be interpreted as the ADMM Algorithm 3 with adaptive vector-valued step-sizes $\boldsymbol{\tau}_r^t$ and $\boldsymbol{\tau}_p^t$ and a particular choice of penalty $Q_x(\cdot)$. Although the convergence of a simpler variant of this algorithm, with non-adaptive step-sizes and convex $f_x(\cdot)$ and $f_z(\cdot)$, was studied in [10]–[12], the results there do not directly apply to the more general GAMP algorithm. However, the second part of the Theorem shows at least that, if the algorithm converges, its fixed points will be critical points of the constrained optimization (1). The full paper [20] also shows that the quadratic term $\boldsymbol{\tau}_x$ can be interpreted as an approximate diagonal to the inverse Hessian.

IV. Fixed-Points of Sum-Product GAMP

The characterization of the fixed points of the sum-product GAMP algorithm is somewhat more complicated to describe, and requires a certain *variational* interpretation – a common framework for understanding sum-product loopy BP [21]. For any fixed observation $\mathbf{y}$, the density function $p_{\mathbf{x},\mathbf{z}|\mathbf{y}}(\cdot,\cdot|\mathbf{y})$ in (3) must minimize

$$p_{\mathbf{x},\mathbf{z}|\mathbf{y}} = \arg\min_b D(b||p_{\mathbf{x},\mathbf{z}|\mathbf{y}}), \tag{11}$$

where the minimization is over all density functions $b(\mathbf{x},\mathbf{z})$ with support restricted to $\mathbf{z} = \mathbf{A}\mathbf{x}$ and $D(b||p_{\mathbf{x},\mathbf{z}|\mathbf{y}})$ is the Kullback-Leibler (KL) divergence. Now, let $p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})$ and $p_{\mathbf{z}|\mathbf{y}}(\mathbf{z}|\mathbf{y})$ be the marginal densities for the posterior $p_{\mathbf{x},\mathbf{z}|\mathbf{y}}$. Using the relationship $b(\mathbf{x},\mathbf{z}) = b_x(\mathbf{x})\mathbb{1}_{\{\mathbf{z}=\mathbf{A}\mathbf{x}\}}$ and the separable nature of $F(\mathbf{x},\mathbf{z})$ in (2), it can be verified that the minimization (11) implies that the marginal densities over $\mathbf{x}$ and $\mathbf{z}$ are solutions to the optimization

$$(p_{\mathbf{x}|\mathbf{y}}, p_{\mathbf{z}|\mathbf{y}}) = \arg\min_{b_x,b_z} J_{\mathrm{KL}}(b_x,b_z) \text{ s.t. } b_z = T_{\mathbf{A}} b_x, \tag{12}$$

where the minimization is over density functions $b_x(\mathbf{x})$ and $b_z(\mathbf{z})$ and $J_{\mathrm{KL}}(b_x,b_z)$ is the functional

$$J_{\mathrm{KL}}(b_x,b_z) := D(b_x||e^{-f_x}) + D(b_z||e^{-f_z}) + H(b_z). \tag{13}$$

In (12), we have used the notation $b_z = T_{\mathbf{A}} b_x$ to indicate that $b_z(\mathbf{z})$ is the density for a random vector $\mathbf{z} = \mathbf{A}\mathbf{x}$ with $\mathbf{x} \sim b_x(\mathbf{x})$. Thus, $p_{\mathbf{z}|\mathbf{y}} = T_{\mathbf{A}} p_{\mathbf{x}|\mathbf{y}}$. Note that we are treating $\mathbf{A}$ as deterministic.

Our next main result, Theorem 2 below, will show that the fixed points of the sum-product GAMP algorithm can be interpreted as critical points of the optimization (12), but with three key approximations: First, similar to what is known as a *mean-field* approximation, the optimization is performed only over factorizable density functions of the form

$$b_x(\mathbf{x}) = \prod_{j=1}^n b_{x_j}(x_j), \quad b_z(\mathbf{z}) = \prod_{i=1}^m b_{z_i}(z_i). \tag{14}$$

Secondly, the objective function J_{KL} from (13) is replaced by

$$\begin{aligned} J_{\mathrm{SP}}(b_x,b_z,\overline{\boldsymbol{\tau}}_p) \;:=\; & D(b_x||e^{-f_x}) + D(b_z||e^{-f_z}) \\ & + H_{\mathrm{gauss}}(b_z,\overline{\boldsymbol{\tau}}_p) \end{aligned} \tag{15}$$

where $\overline{\boldsymbol{\tau}}_p$ is a positive vector and $H_{\mathrm{gauss}}(b_z,\overline{\boldsymbol{\tau}}_p)$ is the following Gaussian upper bound on the entropy $H(b_z)$:

$$H_{\mathrm{gauss}}(b_z,\overline{\boldsymbol{\tau}}_p) := \sum_{i=1}^m \left[\frac{1}{2\overline{\tau}_{p_i}} \mathrm{var}(z_i|b_{z_i}) + \frac{1}{2}\log(2\pi\overline{\tau}_{p_i}) \right]. \tag{16}$$

The third and final approximation is that the constraint $b_z = T_{\mathbf{A}} b_x$ is replaced by the weaker *moment matching* constraint pair $\mathbb{E}(\mathbf{z}|b_z) = \mathbf{A}\mathbb{E}(\mathbf{x}|b_x)$ and $\overline{\boldsymbol{\tau}}_p = \mathbf{S}\mathrm{var}(\mathbf{x}|b_x)$, where $\mathbf{S}$ is given in line 4 of Algorithm 1. The resulting optimization is

$$(\widehat{b}_x,\widehat{b}_z,\boldsymbol{\tau}_p) = \arg\min_{b_x,b_z,\overline{\boldsymbol{\tau}}_p} J_{\mathrm{SP}}(b_x,b_z,\overline{\boldsymbol{\tau}}_p) \tag{17a}$$

$$\text{s.t. } \mathbb{E}(\mathbf{z}|b_z) = \mathbf{A}\mathbb{E}(\mathbf{x}|b_x), \quad \overline{\boldsymbol{\tau}}_p = \mathbf{S}\mathrm{var}(\mathbf{x}|b_x). \tag{17b}$$

The full paper [20] provides further discussion on this approximate optimization. Corresponding to the approximate optimization (17), define

$$F_{\mathrm{SP}}(\overline{\mathbf{x}},\overline{\mathbf{z}},\overline{\boldsymbol{\tau}}_x,\overline{\boldsymbol{\tau}}_p) = F^x_{\mathrm{SP}}(\overline{\mathbf{x}},\overline{\boldsymbol{\tau}}_x) + F^z_{\mathrm{SP}}(\overline{\mathbf{z}},\overline{\boldsymbol{\tau}}_p). \tag{18}$$

where the terms on the right-hand side are the constrained optima

$$\begin{aligned} F^x_{\mathrm{SP}}(\overline{\mathbf{x}},\overline{\boldsymbol{\tau}}_x) \;:=\; & \min_{b_x} D(b_x||e^{-f_x}) \\ & \text{s.t. } \mathbb{E}(\mathbf{x}|b_x) = \overline{\mathbf{x}}, \quad \mathrm{var}(\mathbf{x}|b_x) = \overline{\boldsymbol{\tau}}_x \end{aligned} \tag{19a}$$

$$\begin{aligned} F^z_{\mathrm{SP}}(\overline{\mathbf{z}},\overline{\boldsymbol{\tau}}_p) \;:=\; & \min_{b_z} D(b_z||e^{-f_z}) + H_{\mathrm{gauss}}(b_z,\overline{\boldsymbol{\tau}}_p) \\ & \text{s.t. } \mathbb{E}(\mathbf{z}|b_z) = \overline{\mathbf{z}}. \end{aligned} \tag{19b}$$

We can then replace the optimization over densities in (17) with the optimization

$$(\widehat{\mathbf{x}},\widehat{\mathbf{z}},\boldsymbol{\tau}_x,\boldsymbol{\tau}_p) \;=\; \arg\min_{\overline{\mathbf{x}},\overline{\mathbf{z}},\overline{\boldsymbol{\tau}}_x,\overline{\boldsymbol{\tau}}_p} F_{\mathrm{SP}}(\overline{\mathbf{x}},\overline{\mathbf{z}},\overline{\boldsymbol{\tau}}_x,\overline{\boldsymbol{\tau}}_p) \tag{20a}$$

$$\text{s.t. } \overline{\mathbf{z}} = \mathbf{A}\overline{\mathbf{x}}, \;\; \overline{\boldsymbol{\tau}}_p = \mathbf{S}\overline{\boldsymbol{\tau}}_x. \tag{20b}$$

See the full paper [20] for more details. Corresponding to (20), define the Lagrangian

$$L_{\mathrm{SP}}(\overline{\mathbf{x}},\overline{\mathbf{z}},\overline{\boldsymbol{\tau}}_x,\overline{\boldsymbol{\tau}}_p,\overline{\mathbf{s}}) = F_{\mathrm{SP}}(\overline{\mathbf{x}},\overline{\mathbf{z}},\overline{\boldsymbol{\tau}}_x,\overline{\boldsymbol{\tau}}_p) + \overline{\mathbf{s}}^T(\overline{\mathbf{z}} - \mathbf{A}\overline{\mathbf{x}}), \tag{21}$$

where $\overline{\mathbf{s}}$ represents a vector of dual parameters. We can now state the main result.

Theorem 2: Consider the outputs of the sum-product GAMP version of Algorithm 1. Then, the updates for $\mathbf{x}^t$ and $\boldsymbol{\tau}_x^t$ are equivalent to

$$\begin{aligned} (\mathbf{x}^{t+1},\boldsymbol{\tau}_x^{t+1}) \;=\; & \arg\min_{\overline{\mathbf{x}},\overline{\boldsymbol{\tau}}_x} \Big[L_{\mathrm{SP}}(\overline{\mathbf{x}},\mathbf{z}^t,\overline{\boldsymbol{\tau}}_x,\boldsymbol{\tau}_p^t,\mathbf{s}^t) \\ & + \frac{1}{2}(\boldsymbol{\tau}_s^t)^T\mathbf{S}\overline{\boldsymbol{\tau}}_x + \frac{1}{2}\|\overline{\mathbf{x}} - \mathbf{x}^t\|^2_{\boldsymbol{\tau}_r^t} \Big]. \end{aligned} \tag{22}$$

where $L_{\mathrm{SP}}(\mathbf{x},\mathbf{z},\mathbf{s})$ is the Lagrangian in (7). In addition, the updates preceding (22) that yield $\boldsymbol{\tau}_p^t$, $\mathbf{z}^t$, $\mathbf{s}^t$, and $\boldsymbol{\tau}_s^t$ are

equivalent to

$$\boldsymbol{\tau}_p^t = \mathbf{S}\boldsymbol{\tau}_x^t \tag{23a}$$

$$\mathbf{z}^t = \arg\min_{\overline{\mathbf{z}}}\Big[L_{\rm SP}(\mathbf{x}^t,\overline{\mathbf{z}},\boldsymbol{\tau}_x^t,\boldsymbol{\tau}_p^t,\mathbf{s}^{t-1}) + \frac{1}{2}\|\overline{\mathbf{z}}-\mathbf{A}\mathbf{x}^t\|^2_{\boldsymbol{\tau}_p^t}\Big] \tag{23b}$$

$$\mathbf{s}^t = \mathbf{s}^{t-1} + \frac{1}{\boldsymbol{\tau}_p^t}(\mathbf{z}^t - \mathbf{A}\mathbf{x}^t) \tag{23c}$$

$$\boldsymbol{\tau}_s^t = 2\frac{\partial}{\partial\boldsymbol{\tau}_p}L_{\rm SP}(\mathbf{x}^t,\mathbf{z}^t,\boldsymbol{\tau}_x^t,\boldsymbol{\tau}_p^t,\mathbf{s}^t). \tag{23d}$$

Moreover, any fixed point of the sum-product GAMP algorithm is a critical point of the Lagrangian (21). In addition, the density functions for the minimization in (19) are given by

$$\widehat{b}^x(\mathbf{x}) = p(\mathbf{x}|\mathbf{r},\boldsymbol{\tau}_r), \quad \widehat{b}^z(\mathbf{z}) = p(\mathbf{z}|\mathbf{p},\boldsymbol{\tau}_p), \tag{24}$$

where $p(\mathbf{x}|\mathbf{r},\boldsymbol{\tau}_r)$ and $p(\mathbf{z}|\mathbf{p},\boldsymbol{\tau}_p)$ are given by (5).

Theorem 2 shows a relation between sum-product GAMP and both the ISTA and ADMM methods described earlier. Specifically, define the variables

$$\mathbf{u} := (\overline{\mathbf{x}},\overline{\boldsymbol{\tau}}_x), \quad \mathbf{v} := (\overline{\mathbf{z}},\overline{\boldsymbol{\tau}}_p).$$

Due to the separable structure of the objective function (19), the optimization (20) can be regarded as minimizing a separable function $F^x_{\rm SP}(\mathbf{u}) + F^z_{\rm SP}(\mathbf{v})$ with linear constraints (20b) between $\mathbf{u}$ and $\mathbf{v}$. In this context, the $\mathbf{x}$ and $\mathbf{z}$ minimizations in (22) and (23b) follow the format of the ADMM minimizations in Algorithm 3 for certain choices of the auxiliary functions. On the other hand, the optimization over $\boldsymbol{\tau}_x$ and $\boldsymbol{\tau}_p$ components follow the gradient-based method in the generalized ISTA method in Algorithm 2. So, the sum-product GAMP algorithm can be seen as a hybrid of the ISTA and ADMM methods for the optimization (20), which is equivalent to the variational optimization (17).

Unfortunately, this hybrid ISTA-ADMM method is non-standard and there is no existing convergence theory on the algorithm. However, Theorem 2 at least shows that if the sum-product GAMP algorithm converges, its fixed points correspond to critical points of optimization (20).

Conclusions

Although AMP methods admit precise analyses in the context of large random transform matrices $\mathbf{A}$, their behavior for general matrices is less well-understood. This limitation is unfortunate since many transforms arising in practical problems such as imaging and regression are not well-modeled as realizations of large random matrices. To help overcome these limitations, this paper draws connections between AMP and certain variants of standard optimization methods that employ adaptive vector-valued step-sizes. These connections enable a precise characterization of the fixed-points of both max-sum and sum-product GAMP for the case of arbitrary transform matrices $\mathbf{A}$. The convergence of AMP methods for general $\mathbf{A}$ is, however, still not fully understood. Simulations (not shown here) have indicated, for example, that under general choices of $\mathbf{A}$, AMP may diverge. We hope that the connections between AMP and standard optimization methods provided here help to better understand, and even improve, AMP convergence with general matrices.

References

[1] P. McCullagh and J. A. Nelder, *Generalized linear models*, 2nd ed. Chapman & Hall, 1989.

[2] Y. C. Eldar and G. Kutyniok, *Compressed Sensing: Theory and Applications*. New York: Cambridge Univ. Press, 2012.

[3] A. Chambolle, R. A. DeVore, N. Y. Lee, and B. J. Lucier, "Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage," *IEEE Trans. Image Process.*, vol. 7, no. 3, pp. 319–335, Mar. 1998.

[4] I. Daubechies, M. Defrise, and C. D. Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Commun. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413–1457, Nov. 2004.

[5] S. J. Wright, R. D. Nowak, and M. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Trans. Signal Process.*, vol. 57, no. 7, pp. 2479–2493, Jul. 2009.

[6] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problem," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[7] Y. E. Nesterov, "Gradient methods for minimizing composite objective function," *CORE Report*, 2007.

[8] J. Bioucas-Dias and M. Figueiredo, "A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration," *IEEE Trans. Image Process.*, vol. 16, no. 12, pp. 2992 – 3004, Dec. 2007.

[9] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, 2010.

[10] J. E. Esser, "Primal dual algorithms for convex models and applications to image restoration, registration and nonlocal inpainting," Ph.D. dissertation, The University of California, Los Angeles, 2010.

[11] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imaging Vis.*, vol. 40, pp. 120–145, 2011.

[12] B. He and X. Yuan, "Convergence analysis of primal-dual algorithms for a saddle-point problem: From contraction perspective," *SIAM J. Imaging Sci.*, vol. 5, no. 1, pp. 119–149, 2012.

[13] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci.*, vol. 106, no. 45, pp. 18 914–18 919, Nov. 2009.

[14] ——, "Message passing algorithms for compressed sensing I: motivation and construction," in *Proc. Info. Theory Workshop*, Jan. 2010.

[15] ——, "Message passing algorithms for compressed sensing II: analysis and validation," in *Proc. Info. Theory Workshop*, Jan. 2010.

[16] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Trans. Inform. Theory*, vol. 57, no. 2, pp. 764–785, Feb. 2011.

[17] S. Rangan, "Estimation with random linear mixing, belief propagation and compressed sensing," in *Proc. Conf. on Inform. Sci. & Sys.*, Princeton, NJ, Mar. 2010, pp. 1–6.

[18] ——, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. IEEE Int. Symp. Inform. Theory*, Saint Petersburg, Russia, Jul.–Aug. 2011, pp. 2174–2178.

[19] A. Javanmard and A. Montanari, "State evolution for general approximate message passing algorithms, with applications to spatial coupling," arXiv:1211.5164 [math.PR]., Nov. 2012.

[20] S. Rangan, P. Schniter, E. Riegler, A. Fletcher, and V. Cevher, "Fixed points of generalized approximate message passing with arbitrary matrices," arXiv preprint, Jan. 2013.

[21] M. Wainwright and M. Jordan, "Graphical models, exponential families, and variational inference," *Found. Trends Mach. Learn.*, vol. 1, 2008.

[22] J. Boutros and G. Caire, "Iterative multiuser joint decoding: Unified framework and asymptotic analysis," *IEEE Trans. Inform. Theory*, vol. 48, no. 7, pp. 1772–1793, Jul. 2002.

[23] T. P. Minka, "A family of algorithms for approximate Bayesian inference," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 2001.

[24] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal Stat. Soc., Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.